Data Warehouse

Program Studi Sistem Informasi
Fakultas Ilmu Komputer dan Rekayasa

Mardiani, S.Si., M.T.I

DATA MINING

Apa Data Mining?

- Data mining (pencarian pengetahuan dari data)
  - Mengekstrak secara otomatis pola atau pengetahuan yang menarik (tidak sederhana, tersembunyi, tidak diketahui sebelumnya, berpotensi berguna) dari data dalam jumlah sangat besar.

Data Mining adalah usaha penemuan pengetahuan di intelejensia buatan (bidang *machine learning*) atau analisis statistik dengan mencari atau menemukan aturan-aturan, pola-pola dan struktur dari himpunan data yang besar.

Mengapa Data Mining: Banjir Data

- Twitter: 8000an tweet per detik → 600 juta tweet per hari.
- Facebook: 30 milyar item (link, status, note, foto dst) per bulan. 500 juta user menghabiskan 700 milyar menit per bulan di situs FB.
- Indomaret: 4500an gerai, asumsikan 3 transaksi per menit = 12 juta transaksi per hari se Indonesia.
- Kartu kredit visa: berlaku di 200 negara. 10 ribu transaksi per detik → 850 juta transaksi per hari.

Apa Datamining?

- Nama alternatif: Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence dsb

- Keuntungan bagi organisasi yang menerapkan data mining?

Keuntungan Datamining

- Perusahaan fokus ke informasi yg berharga di datawarehouse/ databasenya.
- Meramalkan masa depan → perusahaan dapat mempersiapkan diri

Contoh:

Midwest grocery chain menggunakan DM untuk menganalisisi pola pembelian: saat pria membeli roti di hari Kamis dan Sabtu, mereka juga membeli minuman.

Analisis lebih lanjut: pembeli ini belanja di hari kamis dan sabtu, tapi di hari kamis jumlah item lebih sedikit. Kesimpulan yang diambil: pembeli membeli minuman untuk dihabiskan saat weekend.

Tindak lanjut: menjual minuman dengan harga full di hari Kamis dan Sabtu. Mendekatkan posisi roti dan minuman.

Lanjutan..

CONTOH KASUS :

Jika Anda mempunyai kartu kredit, sudah pasti Anda bakal sering menerima surat berisi brosur penawaran barang atau jasa. Jika Bank pemberi kartu kredit Anda mempunyai 1.000.000 nasabah, dan mengirimkan sebuah (hanya satu) penawaran dengan biaya pengiriman sebesar Rp. 1.000 per buah maka biaya yang dihabiskan adalah Rp. 1 Milyar!! Jika Bank tersebut mengirimkan penawaran sekali sebulan yang berarti 12x dalam setahun maka anggaran yang dikeluarkan per tahunnya adalah Rp. 12 Milyar!! Dari dana Rp. 12 Milyar yang dikeluarkan, berapa persenkah konsumen yang benar-benar membeli? Mungkin hanya 10 %-nya saja. Secara harfiah, berarti 90% dari dana tersebut terbuang sia-sia.

Lanjutan..

- Dari contoh kasus di atas merupakan salah satu persoalan yang dapat diatasi oleh data mining dari sekian banyak potensi permasalahan yang ada. Data mining dapat menambang data transaksi belanja kartu kredit untuk melihat manakah pembeli-pembeli yang memang potensial untuk membeli produk tertentu. Mungkin tidak sampai presisi 10%, tapi bayangkan jika kita dapat menyaring 20% saja, tentunya 80% dana dapat digunakan untuk hal lainnya.

Contoh

- Proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data dengan tujuan untuk dapat memprediksi kelas dari suatu objek yang labelnya tidak diketahui
- **Contoh : Mendeteksi Penipuan**
- **Tujuan : Memprediksi kasus kecurangan transaksi kartu kredit.**
  - **Pendekatan :**
    - **Menggunakan transaksi kartu kredit dan informasi dilihat dari atribut account holder**
      - **Kapan cutomer melakukan pembelian, Dengan cara apa customer membayar, sebarapa sering customer membayar secara tepat waktu, dll**
    - **Beri nama/tanda transaksi yang telah dilaksanakan sebagai transaksi yang curang atau yang baik. Ini sebagai atribut klass (the class attribute.)**
    - **Pelajari model untuk class transaksi**
    - **Gunakan model ini untuk mendeteksi kecurangan dengan meneliti transaksi kartu kredit pada account.**

Contoh Aplikasi

Bank me-mining transaksi customer untuk mengidentifikasi customer yang kemungkinan besar tertarik terhadap produk baru.

Setelah teknik ini digunakan, terjadi peningkatan **20 kali lipat penurunan biaya** dibandingkan dengan cara biasa.

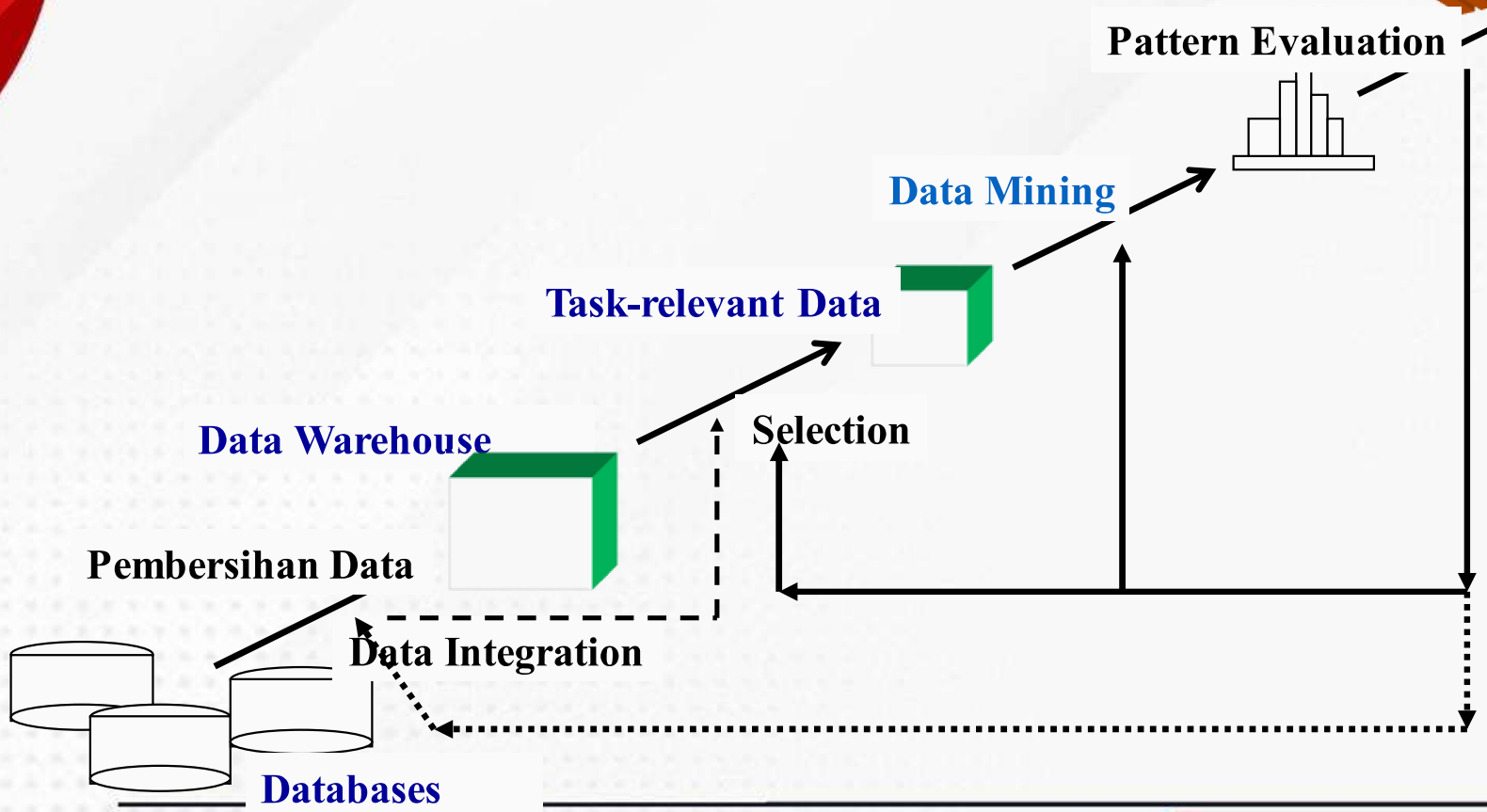
Contoh Aplikasi

Perusahaan transportasi memining data customer untuk mengelompokkan customer yang memiliki nilai tinggi yang perlu diprioritaskan.

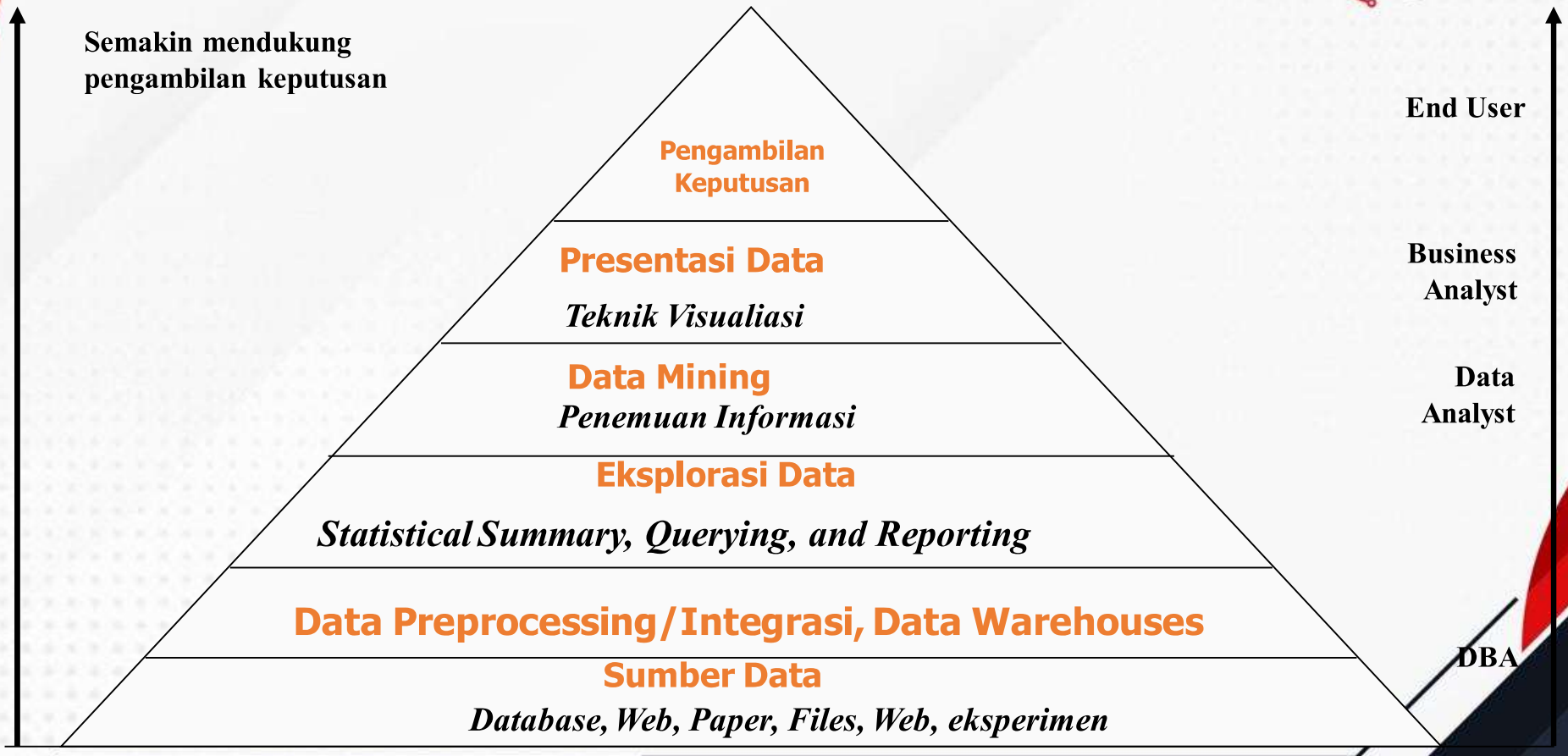
Data Mining pada Industri Retail

- Industri Retail: besarnya data penjualan, sejarah belanja pelanggan, dan lain-lain
- Aplikasi dari Retail data mining
  - Mengidentifikasi perilaku pembelian pelanggan
  - Menentukan kecenderungan pola belanja pelanggan
  - Meningkatkan mutu dari layanan pelanggan
  - Mencapai kepuasan pelanggan
  - Tingkatkan perbandingan konsumsi barang-barang
  - Mendisain keefektifan distribusi dan transportasi barang

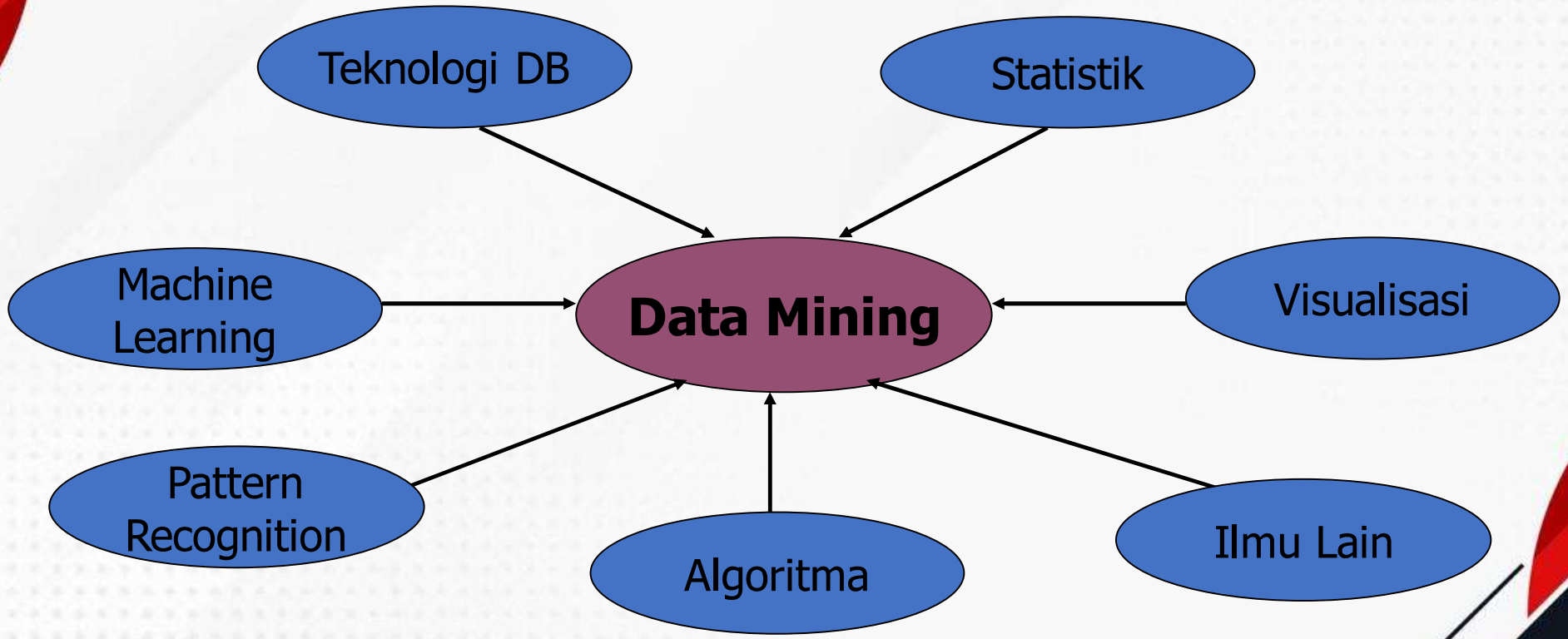
Proses Datamining



Data Mining dan Business Intelligence



Data Mining: Multi Disiplin Ilmu



Mengapa tidak analisis data biasa?

- Jumlah data yang sangat besar
  - Algoritma harus scalable untuk menangani data yang sangat besar (tera)
- Dimensi yang sangat besar: ribuan field
- Data Kompleks
  - Aliran data dan sensor
  - Data terstruktur, graph, social network, multi-linked data
  - Database dari berbagai sumber, database lama
  - Spasial (peta), multimedia, text, web
  - Software Simulator

Data Mining dari berbagai sudut pandang

- **Data**
  - Relational, data warehouse, web, transactional, stream, OO, spacial, text, multimedia
- **Pengetahuan yang akan ditambang**
  - Karakterisitik, diskriminasi, asosiasi, klasifikasi, clustering, trend, outlier
- **Teknik**
  - Database, OLAP, machine learning, statistik, visualiasi
- **Penerapan**
  - Retail, telekomunikasi, banking, analisis kejahatan, bio-data mining, saham, text mining, web mining

Model dalam Data Mining

- **Verification Model**
  - Model ini menggunakan (hypothesis) dari pengguna, dan melakukan test terhadap perkiraan yang diambil sebelumnya dengan menggunakan data-data yang ada.
  - Model *verifikasi* menggunakan pendekatan *top down* dengan mengambil hipotesa dari user dan memeriksa validitasnya dengan data sehingga bisa dibuktikan kebenaran hipotesa tersebut.

Model dalam Data Mining

- **Discovery Model**
  - Pada *directed knowledge discovery*, data mining akan mencoba mencari penjelasan nilai target field tertentu (seperti penghasilan, respons, usia, dan lain-lain) terhadap field-field yang lain.
  - Pada *undirected knowledge discovery* tidak ada target field karena komputer akan mencari pola yang ada pada data. Jadi *undirected knowledge discovery* digunakan untuk mengenali hubungan/relasi yang ada pada data sedangkan directed discovery akan menjelaskan hubungan/relasi tersebut.

Data Mining: Data apa saja?

- **Database Tradisional**
  - Relational database, data warehouse, transactional database
- **Advanced Database**
  - Data streams dan data sensor
  - Time-series data, temporal data, sequence data (incl. bio-sequences)
  - Structure data, graphs, social networks and multi-linked data
  - Object-relational databases
  - Heterogeneous databases dan legacy databases
  - Spatial data dan spatiotemporal data
  - Multimedia database
  - Text databases
  - World-Wide Web

Top-10 Algorithm di ICDM'06

- **#1: C4.5 (61 votes)**
- **#2: K-Means (60 votes)**
- **#3: SVM (Support Vector Machine)(58 votes)**
- **#4: Apriori (52 votes)**
- **#5: EM (Expectation Maximization) (48 votes)**
- **#6: PageRank (46 votes)**
- **#7: AdaBoost (45 votes)**
- **#7: kNN (45 votes)**
- **#7: Naive Bayes (45 votes)**
- **#10: CART (Classification and Regression Tree)(34 votes)**

Aplikasi Data Mining

Pemasaran/ Penyewaan

- Identifikasi pola pembayaran pelanggan
- Menemukan asosiasi diantara karakteristik demografik pelanggan
- Analisis keranjang pemasaran

Perbankan

- Mendeteksi pola penyalahgunaan kartu kredit
- Identifikasi pelanggan yang loyal
- Mendeteksi kartu kredit yang dihabiskan oleh kelompok pelanggan

Asuransi & Pelayanan Kesehatan

- Analisis dari klaim
- Memprediksi pelanggan yang akan membeli polis baru
- Identifikasi pola perilaku pelanggan yang berbahaya

Aplikasi Data Mining

- Analisa Perusahaan dan Manajemen Resiko
  - Perencanaan Keuangan dan Evaluasi Aset
  - Perencanaan Sumber Daya (Resource Planning)
  - Persaingan (competition) → Competitive Intelligence
- Telekomunication
  - menerapkan data mining untuk melihat dari jutaan transaksi yang masuk, transaksi mana saja yang masih harus ditangani secara manual (dilayani oleh orang).

Fungsi Data Mining

1. Fungsi Minor atau fungsi tambahan
    * Deskription (deskripsi)
    * Estimation (estimasi)
    * Prediction (prediksi)
2. Fungsi Mayor atau fungsi utama
    * Classification (klasifikasi)
    * Clustering (pengelompokan)
    * Association (asosiasi)

Fungsi Minor

•Deskripsi

Terkadang peneliti dan analis secara sederhana ingin mencoba mencari cara untuk menggambarkan pola dan kecendrungan yang terdapat dalam data yang dimiliki.

•Estimasi

Estimasi hampir sama dengan klasifikasi, kecuali variabel target estimasi lebih ke arah numerik daripada ke arah kategori. Model dibangun menggunakan record lengkap yang menyediakan nilai dari variabel target sebagai nilai prediksi.

•Prediksi

Prediksi hampir sama dengan klasifikasi dan estimasi, kecuali bahwa dalam prediksi nilai dari hasil akan ada di masa mendatang.

Fungsi Mayor

- Klasifikasi

Dalam klasifikasi terdapat target variabel kategori, misal penggolongan pendapatan dapat dipisahkan dalam tiga kategori, yaitu tinggi, sedang dan rendah.

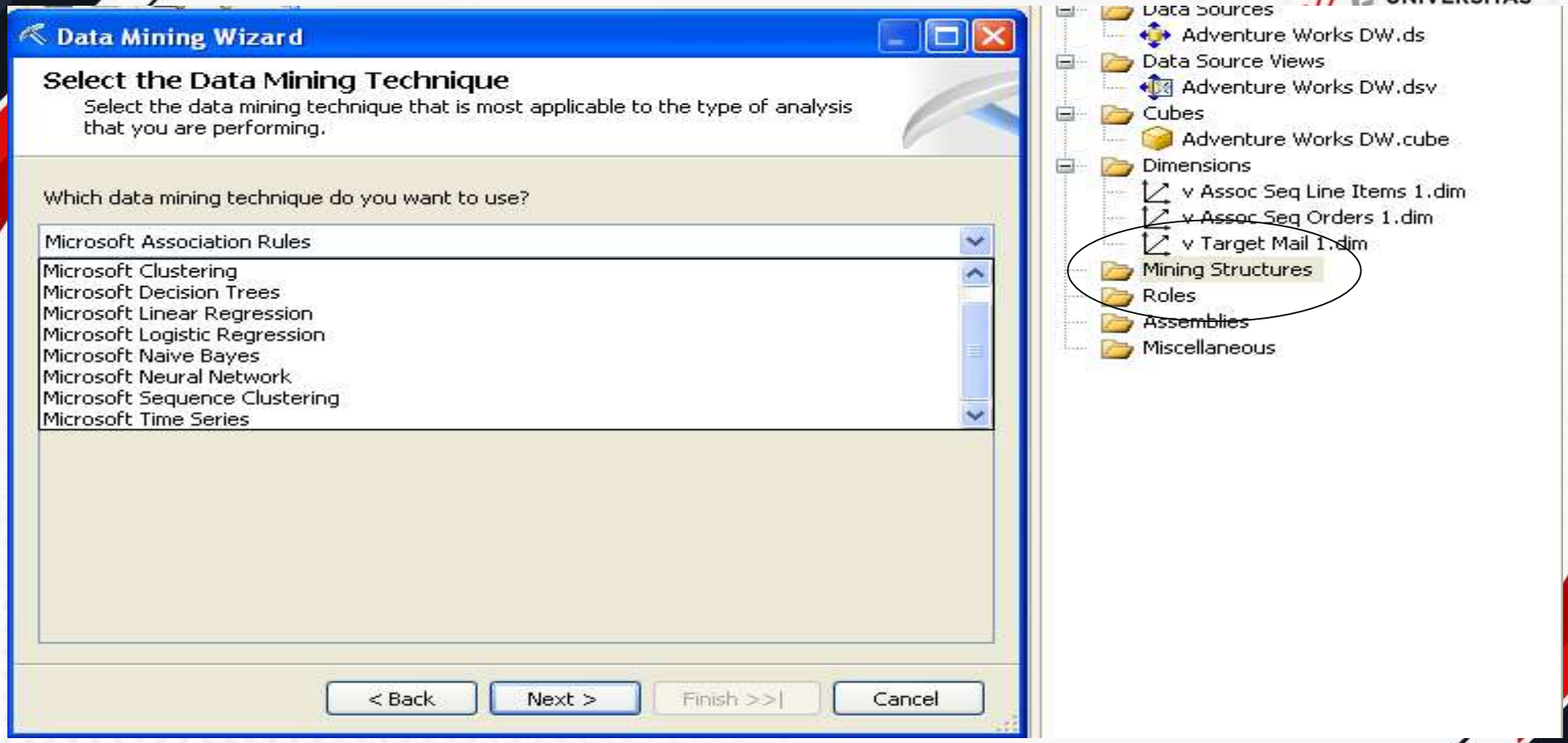
- Pengklusteran

Pengklusteran merupakan pengelompokan record, pengamatan, atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan.

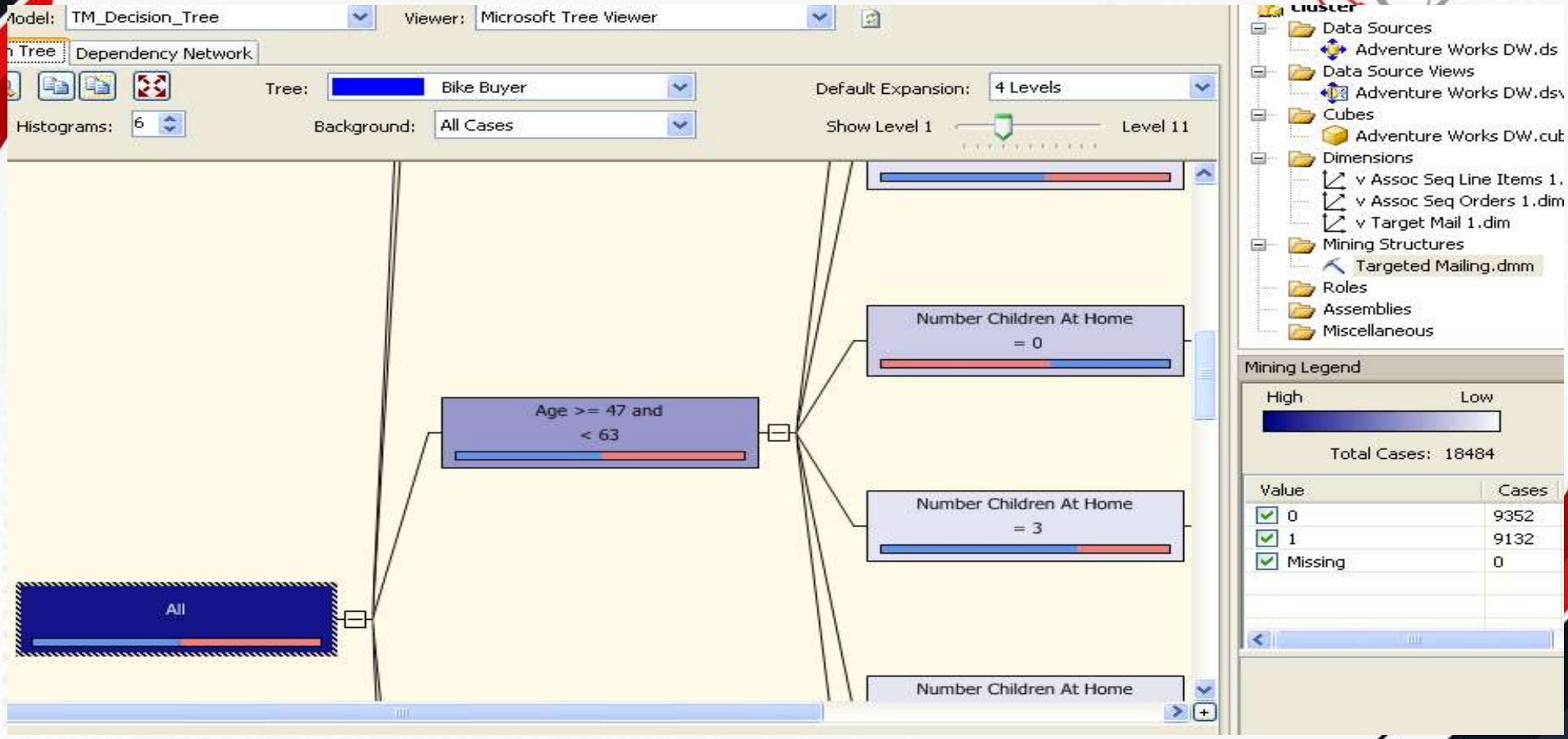
- Asosiasi

Tugas asosiasi *data mining* adalah menemukan atribut yang muncul dalam satu waktu. Dalam dunia bisnis lebih umum disebut analisis keranjang belanja

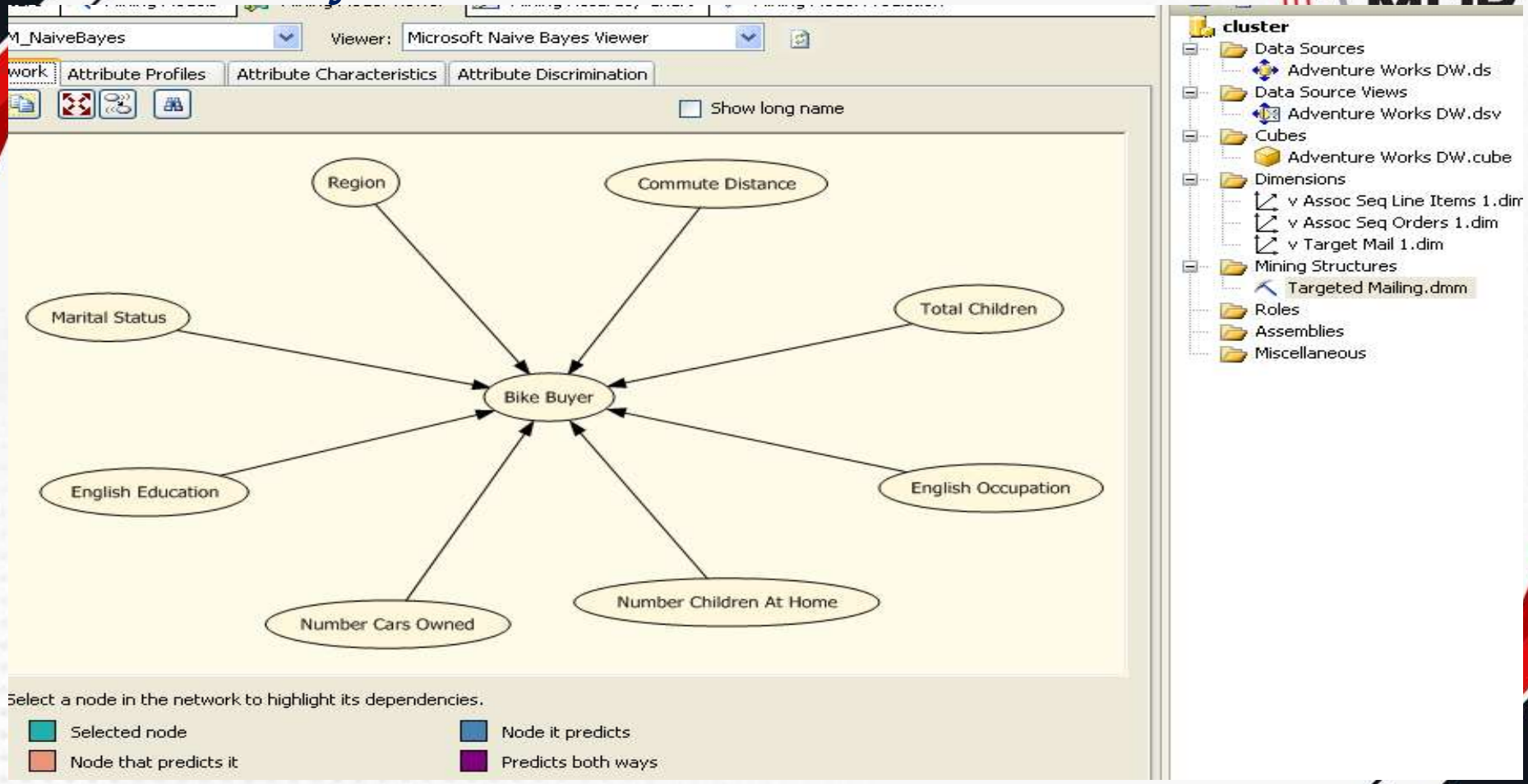
Data Mining Menggunakan Business Intelligence


Data Mining Wizard
Select the Data Mining Technique
Select the data mining technique that is most applicable to the type of analysis that you are performing.
Which data mining technique do you want to use?
Microsoft Association Rules
Microsoft Clustering
Microsoft Decision Trees
Microsoft Linear Regression
Microsoft Logistic Regression
Microsoft Naive Bayes
Microsoft Neural Network
Microsoft Sequence Clustering
Microsoft Time Series
< Back
Next >
Finish >>|
Cancel
Data Sources
Adventure Works DW.ds
Data Source Views
Adventure Works DW.dsv
Cubes
Adventure Works DW.cube
Dimensions
v Assoc Seq Line Items 1.dim
v Assoc Seq Orders 1.dim
v Target Mail 1.dim
Mining Structures
Roles
Assemblies
Miscellaneous

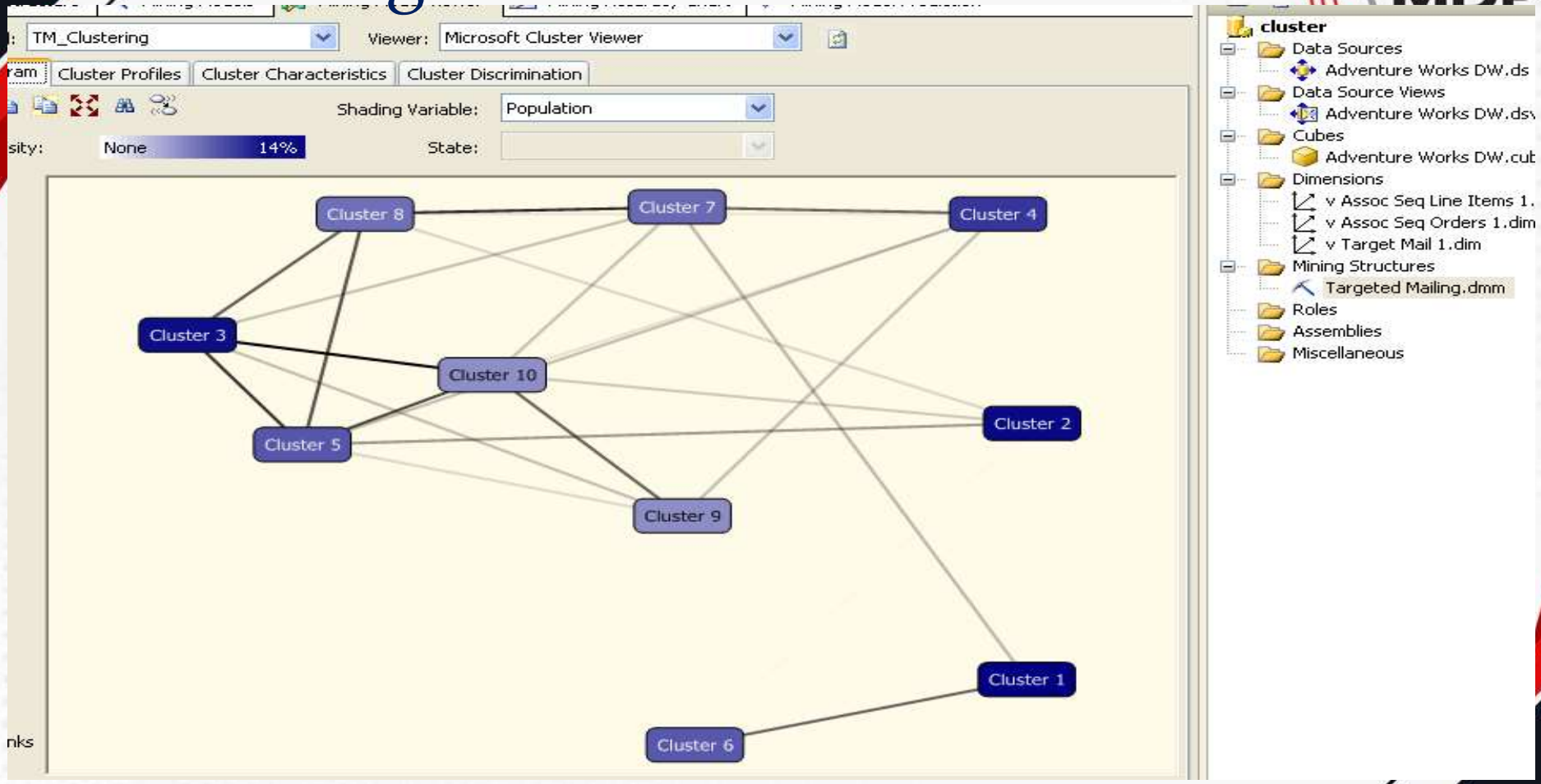
Decision Tree



Naive Bayes


M_NaiveBayes
Viewer:
Microsoft Naive Bayes Viewer
work
Attribute Profiles
Attribute Characteristics
Attribute Discrimination
Show long name
Region
Commute Distance
Marital Status
Total Children
Bike Buyer
English Education
English Occupation
Number Children At Home
Number Cars Owned
Select a node in the network to highlight its dependencies.
Selected node
Node it predicts
Node that predicts it
Predicts both ways
cluster
Data Sources
Adventure Works DW.ds
Data Source Views
Adventure Works DW.dsv
Cubes
Adventure Works DW.cube
Dimensions
v Assoc Seq Line Items 1.dim
v Assoc Seq Orders 1.dim
v Target Mail 1.dim
Mining Structures
Targeted Mailing.dmm
Roles
Assemblies
Miscellaneous

Clustering

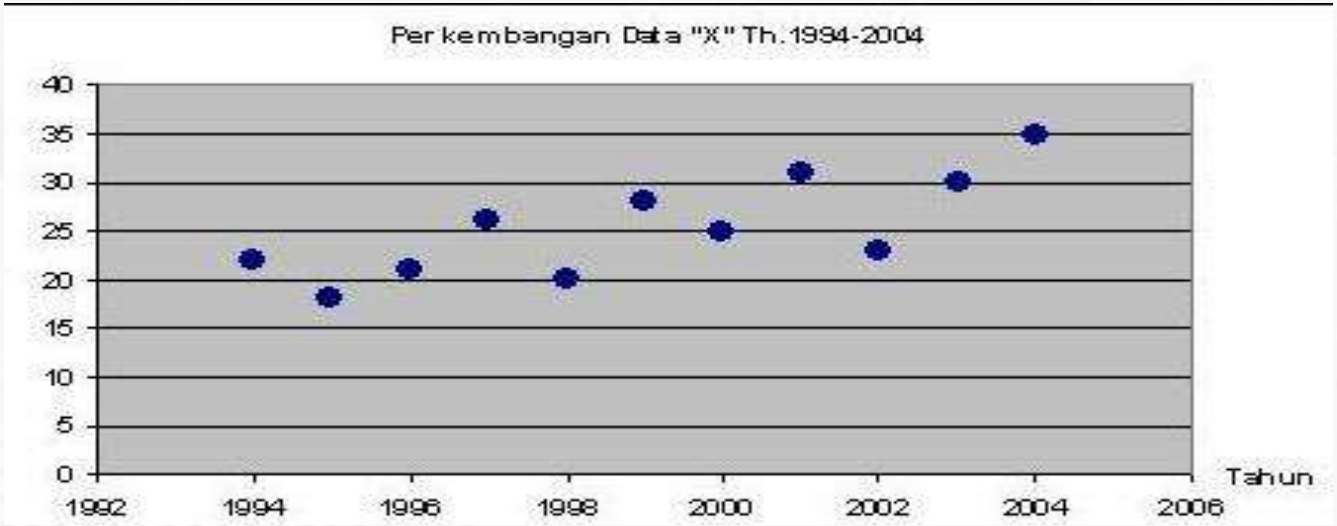

TM_Clustering
Viewer:
Microsoft Cluster Viewer
Cluster Profiles
Cluster Characteristics
Cluster Discrimination
Shading Variable:
Population
None
14%
State:
Cluster 8
Cluster 7
Cluster 4
Cluster 3
Cluster 10
Cluster 2
Cluster 5
Cluster 9
Cluster 1
Cluster 6
cluster
Data Sources
Adventure Works DW.ds
Data Source Views
Cubes
Adventure Works DW.cub
Dimensions
v Assoc Seq Line Items 1.
v Assoc Seq Orders 1.dim
v Target Mail 1.dim
Mining Structures
Targeted Mailing.dmm
Roles
Assemblies
Miscellaneous

FUNGSI MINOR

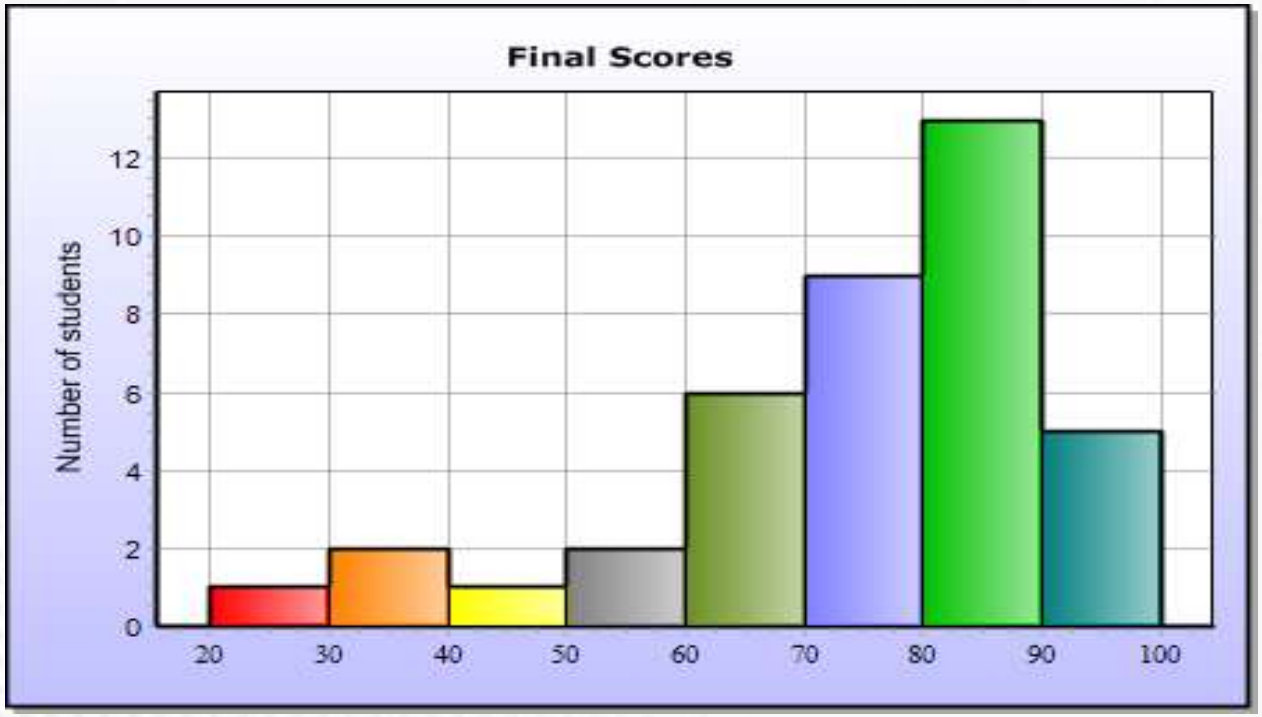
1. Description

- Deskripsi Grafis
  * Diagram Titik
  * Histogram
- Deskripsi Lokasi
  * Rata-rata
  * Median
  * Modus
  * Kuartil, Desil dan Persentil
- Deskripsi Keberagaman
  - Range (rentang)
  - Varians dan Standar Deviasi

Diagram Titik



Histogram


Final Scores
Number of students
0
2
4
6
8
10
12
20
30
40
50
60
70
80
90
100

Rata-rata

- **adalah nilai tunggal yang dianggap dapat mewakili keseluruhan nilai dalam data**

$$\mu = \frac{1}{N} \sum Xi$$

$$\overline{X} = \frac{\sum f_i X_i}{\sum f_i}$$

Median

- **adalah nilai tengah dari data yang ada setelah data diurutkan**

$$Med = X_{\frac{n+1}{2}}$$

$$Med = \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2}$$

$$Med = L_0 + c\left\{\frac{\frac{n}{2} - \left(\sum f_1\right)_0}{f_m}\right\}$$

Modus

- **adalah nilai yang paling sering muncul dalam data**

$$Mod = L_0 + c\left\{\frac{(f_1)_0}{(f_1)_0 + (f_2)_0}\right\}$$

Kuartil, Desil dan Persentil

- **Adalah nilai-nilai yang membagi seperangkat data yang telah terurut menjadi beberapa bagian yang sama**

$$Q_i = L_0 + c\left\{\frac{\frac{in}{4} - \left(\sum f_i\right)_0}{f_q}\right\}$$

$$D_i = L_0 + c\left\{\frac{\frac{in}{10} - \left(\sum f_i\right)_0}{f_d}\right\}$$

$$P_i = L_0 + c\left\{\frac{\frac{in}{100} - \left(\sum f_i\right)_0}{f_p}\right\}$$

Range (rentang)

Nilai Jarak = Nilai Maksimum – Nilai Minimum

Varians dan Standar Deviasi

$$\sigma^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N}$$

2.Estimation

- Rata-rata sampel sebagai estimasi rata-rata populasi
- Varians sampel sebagai estimasi varians populasi
- Standar Deviasi sampel sebagai standar deviasi populasi

3.Prediction

- Regresi Linier Sederhana
- Regresi Linier Berganda

Regresi Linier Sederhana

$$Y' = a + b X$$

a = Y pintasan, (nilai Y' bila X = 0)

b = kemiringan garis regresi

X = nilai tertentu dari variabel bebas

Y'= nilai yang dihitung pada variabel tidak bebas.

$$b = \frac{\sum x_i y_i}{\sum x_i^2} \qquad a = \overline{Y} - b\overline{X}$$

Regresi Linier Berganda

$$Y' = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k$$

$$b_0 n + b_1 \sum X_1 + b_2 \sum X_2 + \ldots + b_k \sum X_k = \sum Y$$

$$b_0 \sum X_1 + b_1 \sum X_1^2 + b_2 \sum X_1 X_2 + \ldots + b_k \sum X_1 X_k = \sum X_1 Y$$

$$b_0 \sum X_2 + b_1 \sum X_1 X_2 + b_2 \sum X_2^2 + \ldots + b_k \sum X_2 X_k = \sum X_2 Y$$

$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots$$

$$b_0 \sum X_k + b_1 \sum X_1 X_k + b_2 \sum X_2 X_k + \ldots + b_k \sum X_k^2 = \sum X_k Y$$

UNIVERSITAS MDP

TERIMA KASIH

KAMPUS INOVASI

Mardiani, S.Si., M.T.I